Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 9662–9666, October 1995
Evolution

Calculating the probability of multitaxon evolutionary trees: Bootstrappers Gambit

(Jeffreys' prior/Bayesian/multinomials/parsimony/eocytes)

JAMES A. LAKE

Molecular Biology Institute and Biology Department, University of California, Los Angeles, CA 90095

Communicated by Richard E. Dickerson, University of California, Los Angeles, CA, July 19, 1995

ABSTRACT **The reconstruction of multitaxon trees from molecular sequences is confounded by the variety of algorithms and criteria used to evaluate trees, making it difficult to compare the results of different analyses. A global method of multitaxon phylogenetic reconstruction described here, Bootstrappers Gambit, can be used with any four-taxon algorithm, including distance, maximum likelihood, and parsimony methods. It incorporates a Bayesian–Jeffreys'–bootstrap analysis to provide a uniform probability-based criterion for comparing the results from diverse algorithms. To examine the usefulness of the method, the origin of the eukaryotes has been investigated by the analysis of ribosomal small subunit RNA sequences. Three common algorithms (paralinear distances, Jukes–Cantor distances, and Kimura distances) support the eocyte topology, whereas one (maximum parsimony) supports the archaebacterial topology, suggesting that the eocyte prokaryotes are the closest prokaryotic relatives of the eukaryotes.**

Determining globally optimal, multitaxon phylogenetic trees is computationally intensive because the number of possible trees increases rapidly with increasing taxa. For four taxa, 3 unrooted trees must be compared, whereas for thirteen, 13,749,310,575 must be compared (1). Evaluating multitaxon trees derived by different methods is further complicated by diverse optimality criteria. For example, distance methods frequently search for local minima by using least-squares criteria, whereas parsimony methods minimize the number of nucleotide changes, often using global searches (2). Currently no common basis exists for reconstructing trees by using different algorithms.

Bayesian and likelihood methods assess the probabilities of trees and thereby can provide a common basis for reconstructing trees by using different algorithms. Sinsheimer *et al.* (3) developed a method for calculating the probability of trees derived by evolutionary parsimony, but the calculations are complex for trees with more than five taxa. Felsenstein (4) has thoughtfully proposed that bootstrap replicates (5, 6) might provide a good method of assessing the likelihood function in tree reconstruction. Both groups calculate the probability, $P(\text{tree}_j|S)$, that the jth tree is correct given aligned sequences, S. These are complex calculations. In this paper one calculates something simpler—the probability, $P(H_j^A|S)$, that algorithm A applied to a sequence of infinite length (generated under the same model as S) would yield the jth tree. Under a multinomial model (assuming a Jeffreys' prior probability on the underlying parameters) the integral for calculating $P(H_j^A|S)$ can be estimated by bootstrap replication. *Bootstrappers Gambit*† combines this bootstrap with a multitaxon algorithm for any four-taxon method.

AN EXAMPLE

Bootstrappers Gambit functions by decomposing multiple taxon trees into sets of four taxon statements as illustrated in Fig. 1 for a five-taxon tree. Five aligned sequences at the top of the figure correspond to taxa 1 through 5. Four bootstrap replicates of the original sequences of the five aligned sequences shown at the top of Fig. 1 were taken by sampling with replacement. Maximum parsimony is used to analyze taxa four at a time, using the neighbors—or for distances, the weak neighbors—relationship (7). For four taxa (i, j, k, and l) three trees are possible (the E tree clusters i with j and k with l; the F tree clusters i with k; and the G tree clusters i with l). For example, in the first column of replicate 1 the quartet represented by taxa 1, 2, 3, and 4 (denoted 1234) corresponds to the sequence pattern AAAA. Since this pattern supports no tree, by parsimony, the result is indicated by a blank (-) in the table of quartet values for replicate 1. In the second column the sequences for quartet 1234 are TTCC. Parsimony interprets this pattern as support for the E tree (8) and an *e* is entered in the quartet value table. The most parsimonious four-taxon trees are then chosen by counting *e*s, *f*s, and *g*s at all sequence positions. The four-taxon trees supported at the most positions are entered into the quartet value table. (If two trees tie, then no tree is selected.) For replicate 1 the pattern of winning four-taxon tree values is EEEEE (quartets 1234, 1235, 1245, 1345, and 2345, respectively). This value pattern is uniquely associated with the tree shown next to the pattern. Some quartet value patterns are inconsistent with trees and may support non-tree graphs (7). For example the pattern from replicate 2, GEEFE, fits no tree. Details of Gambit, used to relate value patterns to trees, are described in *Appendix*.

The last step involves calculating the probability of each tree. The conditional probability that a particular tree would be supported with infinite data is given by the number of replicates supporting the tree divided by the total number of replicates supporting trees (see *Appendix*). In the example two trees corresponding to the EEEEE pattern are present and the total number of trees is three, so that the probability of the EEEEE tree is estimated as 2/3 and the probability of the GEFFF tree is 1/3. Better estimates can be provided by taking more replicates.

RESULTS

Computational Times. The speed of Bootstrappers Gambit depends on the internal consistency of the data. It is fast for consistent data sets and slow for sets with little, or no, information content. To illustrate how times depend upon sequence lengths, trees have been calculated from a set of nine rRNA sequences (for taxa see Fig. 2 legend).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: STSV, site-to-site rate variation.

†*Bootstrappers* refers to the Bayesian–Jeffreys'–bootstrap method of estimating probabilities. *Gambit* (or dance) refers to the systematic phylogeny search.

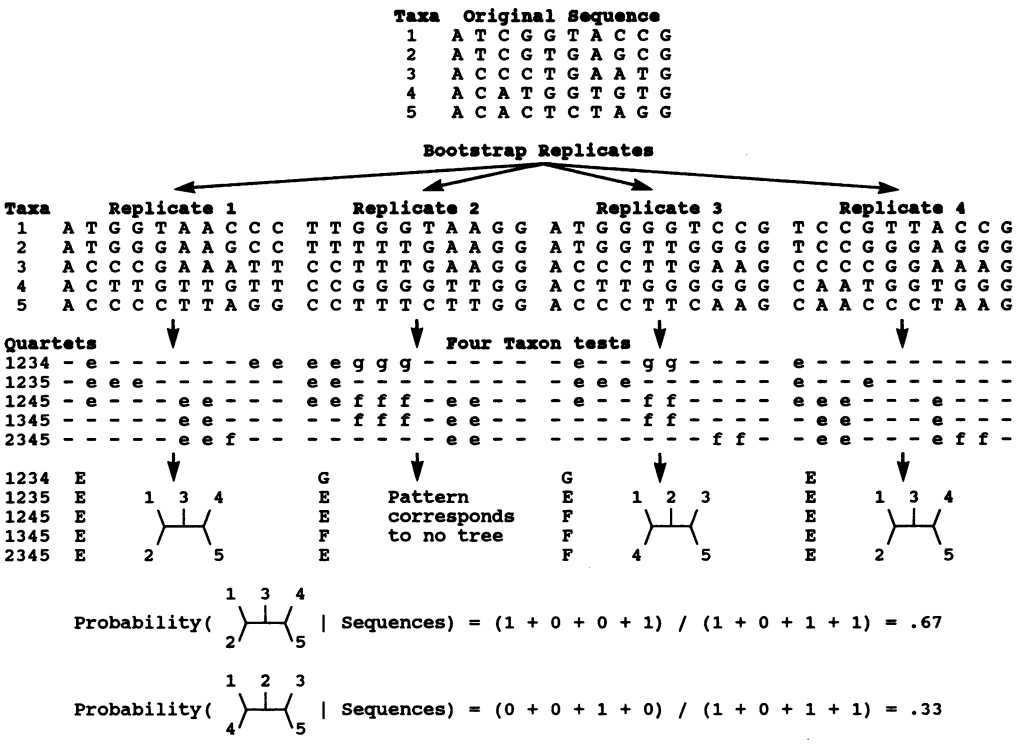

FIG. 1. A five-taxon example of Bootstrappers Gambit. Parsimony is illustrated. Distance methods are similar but use four-point equations to calculate winning four-taxon values.

The mean time to calculate a tree is shown in Fig. 2*A*. For seven taxa, the logarithm of the mean time per tree increases linearly with sequence length, whereas for eight and nine taxa the results are nonlinear, with minima at lengths of 1000 (eight taxa) and 1300 (nine taxa) nucleotides. Times range from less than 1 sec per tree to nearly 1000 sec per tree.

Calculation times can be reduced considerably by relaxing the requirement for 100% nodal consensus (see *Appendix*), but the results may depend on the order of taxon presentation. The comparative time savings that can be obtained with 75% nodal consensus are illustrated in Fig. 2*B*. Times range from 0.5 sec per tree to 10 sec per tree. For internally consistent data using the 100% model, one can analyze up to 9–12 taxa on a personal computer (10), and the 75% consensus model can extend this further (11).

An Example: The Origin of the Eukaryotes. A classical biological problem, determining the origin of the eukaryotes (12), illustrates the usefulness of Bootstrappers Gambit. Because the tree of life spans large evolutionary distances, its reconstruction is strongly affected by unequal rate effects, site-to-site rate variation (STSV), and alignment biases (9, 13–16) which cause incorrect trees to be reconstructed. Hence the origin of eukaryotes is a hard problem with low information content and, potentially, long calculation times.

Classically eukaryotes and prokaryotes have been considered to be two fundamental divisions of life; however, eukaryotes are defined by the *presence* of a positive character, the nucleus, whereas prokaryotes are characterized by the *absence* of a nucleus (17). Assuming the nucleus to be the synapomorphy, prokaryotes may be a heterogenous or paraphyletic group. Two mutually exclusive theories exist for their origin. In one, the eocyte theory, eocytes (hyperthermophilic, sulfur-metabolizing prokaryotes) are the closest prokaryotic relatives of the eukaryotes (13). In the other, the archaebacterial theory, halobacteria, methanogens, and eocytes are all equidistant (in time) from the eukaryotes (18).

To test Bootstrappers Gambit, seven diverse taxa were analyzed (at 100% nodal consensus). These included sequences from a eukaryote, a eubacterium, a methanogen, a halobacterium, *Methanopyrus* (intermediate between halobacteria/methanogens and eocytes/eukaryotes), and two eocyte sequences. Four algorithms, maximum parsimony (8), Jukes–Cantor distances (19), Kimura two-parameter distances (20), and paralinear/log det distances (21–23) were tested. Only the paralinear/log det distances algorithm is insensitive to unequal rate effects, in the absence of STSV (21–23). (All four algorithms are sensitive to STSV and no attempt was made here to correct for this.) The two most probable trees based on our analysis, displayed in Fig. 3, are arbitrarily rooted in the eubacterial branch (24, 25). The eocyte tree is the most probable tree by paralinear distances (69.0%), by Jukes–Cantor distances (80.0%), and by Kimura two-parameter distances (71.0%). Maximum parsimony (84.0%) selected the archaebacterial tree.

DISCUSSION

In principle, the Bayesian–Jeffreys'–bootstrap method can be used with Gambit or with any other multitaxon reconstruction algorithm, including regular parsimony. Since Gambit and regular parsimony can potentially produce different results, it is not clear which method might be better. Gambit parsimony examines only the set of four-taxon marginal distributions of the data, whereas regular parsimony uses complete data and hence may be better. Alternatively, Gambit parsimony may be better because it accepts only tree-like data (7) and, unlike regular parsimony, does not force non-tree-like data to fit a tree. Clearly further research is needed.

The analyses presented here are based on a multinomial model in which a large number of outcomes are considered. For example, for five-taxon parsimony there are 25 informative outcomes, and for four-taxon paralinear distances there may be 256. This means for reasonable sequence lengths, N, one may frequently consider nearly N-variate multinomials. Hence, the Bayesian–Jeffreys'–bootstrap must perform well even if counts of only one or two are observed. This seems to

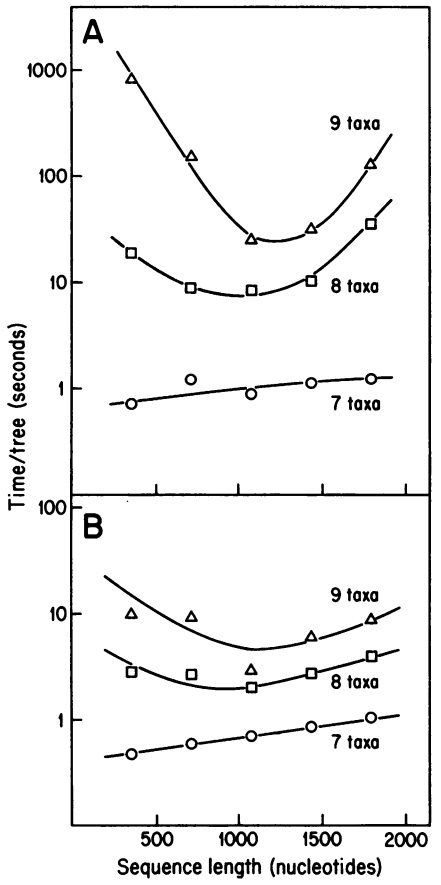

FIG. 2. Computational times for multitaxon trees. Trees were calculated from the following eukaryotes, in the order in which they were entered: *Homo sapiens, Rattus norvegicus, Xenopus laevis, Artemia salina, Saccharomyces cerevisiae, Prorocentrum micans, Euglena gracilis, Zea mays*, and *Glycine max*. Paralinear distances were used for these calculations, but similar results were obtained for other algorithms (data not shown). Shorter sequences were obtained by analyzing only the 5'-most portions (20%, 40%, 60%, and 80%) of the 18S rRNA sequence. Mean calculation times per tree (on a Compaq PC 386 running at 20 MHz) are indicated on the vertical axis. Sequence lengths are indicated on the horizontal axis. Nodal consensus is 100% for *A* and 75% for *B*. While the probabilities of trees are independent of the order of taxon presentation, calculation times depend on the order. The mean times per tree are geometric means calculated in the order listed above and in reverse order. Pairwise alignments of rRNA sequences were performed with the ALIGN program available in the Dayhoff package (for conditions see ref. 9). Star alignments used *Saccharomyces* as the reference.

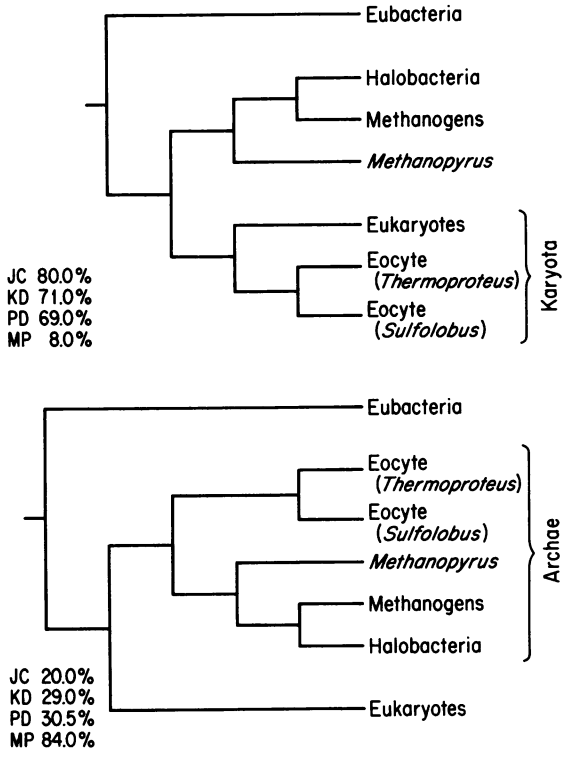

FIG. 3. The two most probable origin of eukaryotes trees obtained by Bootstrappers Gambit. Four four-taxon algorithms were used: Jukes–Cantor (JC), Kimura two-parameter distances (KD), paralinear distances (PD), and maximum parsimony (MP). The probability of each tree is indicated at the lower left of the trees. To reduce alignment artifacts which can cause incorrect trees to be selected, sequences were aligned by using the star alignment (9) with *Thermoproteus tenax* as the reference. Because every bootstrap replicate is not necessarily consistent with a tree, variable numbers of replicates were analyzed. Two hundred trees were calculated for each algorithm. This required the calculation of 2261, 2626, 2460, and 30,358 replicates for the paralinear distance, the two-parameter distance, and the Jukes–Cantor and maximum parsimony algorithms, respectively. Evolutionary parsimony would have required about 666,000 patterns and was stopped for running time considerations before calculations were completed. All trees were calculated using 100% nodal consistency.

be the case. Explicit numerical integrations for trinomials (J.A.L., J. S. Sinsheimer, and R. J. A. Little, unpublished work) indicate that even for the lowest count situations, such as $e = 1, f = 2$, and $g = 3$, probabilities of trees calculated with the bootstrap have a precision of approximately 1.2% when referenced to the direct Bayesian–Jeffreys' calculations.

In the origin of eukaryotes example it is not surprising that alternative trees were obtained. The biases caused by unequal rates, STSV, and alignment order artifactually favor the archaebacterial tree, since the long branches of the eukaryotic and eubacterial taxa attract (9, 14, 15, 21, 26, 27). Nevertheless, three of four of these analyses favored the eocyte tree. Because parsimony (which supports the archaebacterial tree) is quite sensitive to unequal rate effects, whereas paralinear distances (which supports the eocyte tree) is less affected, parsimony's anomalous result may have been due to unequal rate effects.

These results are also consistent with four of five recent studies of elongation factor EF-Tu genes. These include distance matrix analysis (28) (eocyte tree favored), maximum parsimony analysis (26) (eocyte tree favored), maximum likelihood analysis (29) (eocyte tree more likely, but not statistically significant), paralinear distances (21) (eocyte tree favored), and distance matrix results (30) (archaebacterial tree favored). Although STSV was not controlled in either the 18S rRNA or the EF-Tu studies, these studies nevertheless add additional support for the eocyte theory.

Clearly, the availability of a general algorithm for constructing multitaxon trees is an advance. One hopes that in the future the artifacts of STSV and alignment biases will be resolved.

APPENDIX

The Bayesian–Jeffreys'–Bootstrap. The following theorem applies in any situation (biological or otherwise) in which there is multinomial distribution.

THEOREM. *Let* S_k *denote the simplex* $S_k = \{[p_1, \ldots, p_k]; p_j \geq 0, p_1 + \ldots + p_k = 1\}$ *(the set of all possible probability vectors for a* k*-variate multinomial distribution). Given a sample* S *of size* n *from a* k*-variate multinomial distribution, let* $\hat{\pi}$ *be the associated frequency vector in* S_k *whose* j*th component is the proportion of the sample resulting in the* j*th type of outcome. Given a subset* H *of* S_k*, and a multinomial distribution, with a Jeffreys' prior on the underlying probability parameter vector* π *(in* S_k*), the poste-*

rior probability that π *lies in* H *based on a sample* S *of* n *observations drawn from this distribution, is (approximately) the same as the proportion of bootstrap replications* S* *of* S, *for which the associated frequency vector* π^* *lies in* H.

PROOF. Given the sample S, $S = [s_1, \ldots, s_k]$, the probability, $P(S^*|\hat{\pi})$, that a *bootstrap replicate* of S, $S^* = [x_1, \ldots, x_k]$, will contain the jth outcome x_j times is

$$\begin{aligned} P(S^*|\hat{\pi}) &= \frac{n!}{x_1! \ldots x_k!}\left[\frac{s_1}{n}\right]^{x_1} \cdots \left[\frac{s_k}{n}\right]^{x_k} \\ &\approx \frac{\sqrt{n}}{2\pi\sqrt{x_1 \ldots x_k}}\left[\frac{s_1}{x_1}\right]^{x_1} \cdots \left[\frac{s_k}{x_1}\right]^{x_k} \\ &\approx \frac{\sqrt{n}}{2\pi\sqrt{x_1 \ldots x_k}} \\ &\quad \times \exp[-\Delta_1 - (\Delta_1^2/2s_1)] \ldots \exp[-\Delta_k - (\Delta_k^2/2s_k)] \\ &= \frac{\sqrt{n}}{2\pi\sqrt{x_1 \ldots x_k}} \exp(-\Delta_1^2/2s_1) \ldots \exp(-\Delta_k^2/2s_k), \quad [1] \end{aligned}$$

where $\hat{\pi} = (s_1/n, \ldots, s_k/n)$ is the maximum likelihood estimate of π. The second line follows from Stirling's approximation, the third from the binomial expansion of $P(S^*|\hat{\pi})$ about its maximum by substitution of $\Delta_j = x_j - s_j$, and the last from $\Delta_1 + \ldots + \Delta_k = 0$.

Given the sample, the probability of the underlying probabilities, π, can be calculated by using Bayes' equation, specifically $P(\pi|S) \propto P(S|\pi)P_J(\pi)$. To make comparison with Eq. **1** easier the πs are transformed, $\pi_j = x_j/n$. $P_J(\pi)$, the Jeffreys' prior on the distribution of the underlying probabilities, is proportional to $(x_1/n)^{-1/2} \ldots (x_k/n)^{-1/2}$. Hence

$$\begin{aligned} P(\pi|S) &\propto \frac{n!}{s_1! \ldots s_k!}\left[\frac{x_1}{n}\right]^{s_1-1/2} \cdots \left[\frac{x_k}{n}\right]^{s_k-1/2} \\ &\approx \frac{C\sqrt{n}}{2\pi\sqrt{x_1 \ldots x_k}}\left[\frac{x_1}{s_1}\right]^{s_1} \cdots \left[\frac{x_k}{s_k}\right]^{s_k} \\ &\approx \frac{C\sqrt{n}}{2\pi\sqrt{x_1 \ldots x_k}} \\ &\quad \times \exp[+\Delta_1 - (\Delta_1^2/2s_1)] \ldots \exp[+\Delta_k - (\Delta_k^2/2s_k)] \\ &= \frac{C\sqrt{n}}{2\pi\sqrt{x_1 \ldots x_k}} \exp(-\Delta_1^2/2s_1) \ldots \exp(-\Delta_k^2/2s_k) \\ &= C\,P(S^*|\hat{\pi}), \quad [2] \end{aligned}$$

where the steps in the calculation are justified as for Eq. **1** and the constant has the value $C = n^{k/2}/(s_1^{1/2} \ldots s_k^{1/2})$. $P(\pi|S)$ is proportional to $P(S^*|\hat{\pi})$ except that $P(S^*|\hat{\pi})$ is restricted to integer values of x_j, whereas $P(\pi|S)$ is not. Noting that $P(S^*|\hat{\pi})$ is a slowly varying function of x_j, the probability of hypothesis H, a subset of S_k, is

$$P(H|S) \propto \int_{\pi \ni H} P(\pi|S)dp_1 \ldots dp_k \propto \sum_{\pi^* \ni H} P(S^*|\hat{\pi}). \quad [3]$$

Hence $P(H|S)$ is (approximately) proportional to the fraction of bootstrap replicates supporting H.

For this paper, let $H = H_j$ (resp. H^A) be the hypothesis that tree j (resp. *any* tree) is the output of algorithm A applied to infinitely long sequences generated under the same multinomial model as the data S. Then the *conditional* probability $P[H_j^A|H^A]$ is (approximately) the same as the ratio of the number of those bootstrap replicates of S that support tree T_j divided by the number of such replicates that support any tree. Assuming a multinomial model (see refs. 31–33) for four-taxon parsimony analysis, $k = 3$ ($S = [e, f, g]$, corresponding to the counts for the E, F, and G patterns, respectively) and for five-taxon parsimony $k = 25$ (since there are 25 informative patterns), etc. Sites are not assumed independent and identically distributed since one can also classify according to site type.

The Gambit Algorithm. Two examples of Gambit are illustrated in Fig. 4. In the first the sequence value pattern is EEEGG, where the letters refer to the topologies of the quartets 1234, 1235, 1245, 1345, and 2345, respectively. Quartet 1234 has value E and corresponds to the tree (cycle free connected graph) at the top of example 1. This tree contains two internal nodes, N1 and N2, each connected to three adjacent nodes. N1 is connected to two external nodes, representing taxa 1 and 2, and to the internal node connecting taxa 3 and 4 (N2). To add taxon 5 to the tree and position it with respect to node N1, two new quartets, 1235 and 1245 (written as $12\frac{3}{4}5$), are evaluated. Taxon 5 clusters with 1, with 2, or with N2 (3 plus 4) if quartets $12\frac{3}{4}5$ both have value G, F, or E, respectively. If 1235 and 1245 have different values (a logical conflict) then the search is terminated. In example 1, the pattern values $12\frac{3}{4}5$ = E indicate that taxon 5 clusters with N2, shown by a directed edge at the bottom of example 1. Analysis of N2 ($\frac{1}{2}345$ = G) indicates that taxon 5 clusters with taxa 1 plus 2. Thus the node connecting taxon 5 can be introduced only at the *sink* between nodes N1 and N2 in the resulting *intree* (34). In example 2, the quartet pattern resembles that found in example 1, except that the pattern values $12\frac{3}{4}5$ and $\frac{1}{2}345$ are reversed. Since $12\frac{3}{4}5$ = G and $\frac{1}{2}345$ = E, there is logical consistency at both nodes. However, the overall sequence pattern is inconsistent (an arboreal inconsistency, see below) since the analysis of N1 clusters taxa 5 and 1, whereas the

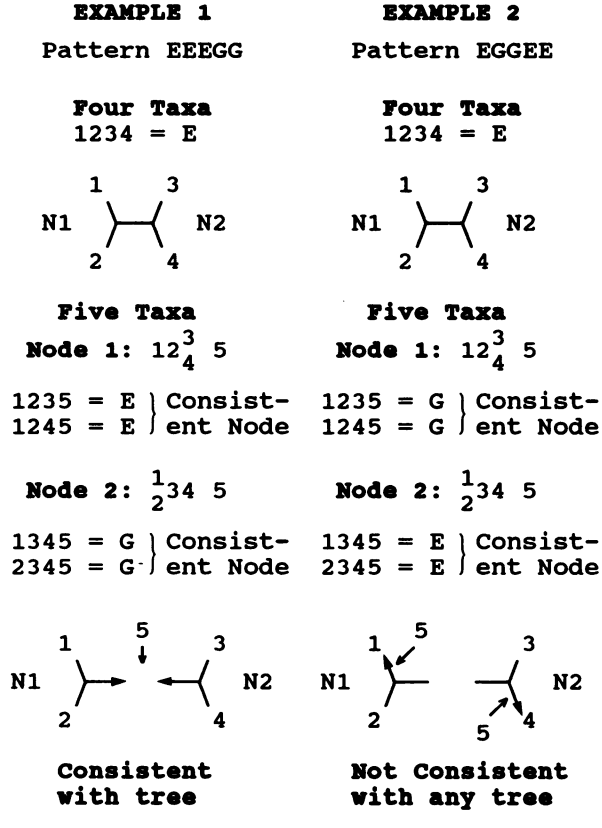

FIG. 4. Two five-taxon examples illustrating the Gambit algorithm.

analysis of N2 clusters taxa 5 and 4 as shown by directed edges at the bottom of Fig. 4.

In general an N-taxon tree can be reconstructed by successively adding taxa and by requiring at each stage that: (*i*) the quartet values calculated about *each* internal node must be the same (nodal consistency), and (*ii*) the acyclic digraph produced by the analysis of quartet values derived from *all* internal nodes is an *intree*, that is, there exists a single insertion site for the taxon being added (arboreal consistency). Condition *i* may be relaxed by requiring only that a given percentage of quartet values be identical. This speeds execution, but for less than 100% consistency solutions are not necessarily independent of the order of taxon presentation.

Bootstrappers Gambit Solutions (for the 100% Consensus Model) Are Global. Because of conditions *i* and *ii*, any tree selected by Gambit will be consistent with all quartets examined during tree construction. Thus if all possible quartets are examined by Gambit [for N taxa, N choose 4, $\binom{N}{4}$, different quartets are possible], then tree selection must be independent of the order of taxon presentation—that is, the solution is global. An inductive proof follows.

First assume that the statement is valid for an $N - 1$ taxon tree [namely, all possible, $\binom{N-1}{4}$, quartets are tested by the Gambit algorithm during reconstruction of the $N - 1$ tree]. Second, note that the set of triples defined by the nodes of the $N - 1$ taxon tree must include all possible, $\binom{N-1}{3}$, triples (since every set of three taxa must intersect at a node in the $N - 1$ taxon tree). Consider now the Gambit extension from $N - 1$ taxa to N taxa. In the extension step a new taxon, taxon N, is added to the tree, so that each triplet node in the $N - 1$ tree, *ijk*, is converted into a quartet, *ijk N*. Since N was not among the taxa included in the $N - 1$ tree, none of these new quartets were in the set of quartets used to reconstruct the $N - 1$ tree. Furthermore, since $\binom{N-1}{4}$ quartets are considered in the $N - 1$ taxon tree and $\binom{N-1}{3}$ new quartets are considered in extending the tree to N taxa, a total of $\binom{N-1}{4} + \binom{N-1}{3} = \binom{N}{4}$ unique quartets will be used to calculate the N-taxon tree. Since there are only $\binom{N}{4}$ possible quartets, all of them must have been evaluated during the construction of the N-taxon tree. Because the assumption is true for four taxa [there is only one, $\binom{4}{4}$, quartet for four taxa], by induction it must also be true for all N-taxon trees (for $N \geq 4$).

I thank my colleagues at the University of California, Los Angeles, and elsewhere for reading early drafts of this manuscript and two dedicated reviewers for their comments. This work was supported by the National Science Foundation and the Sloan Foundation.

1. Cavalli-Sforza, L. L. & Edwards, A. W. F. (1967) *Am. J. Hum. Genet.* **19,** 233–257.
2. Hillis, D. M. & Moritz, C., eds. (1990) *Molecular Systematics* (Sinauer, Sunderland, MA).
3. Sinsheimer, J. S., Lake, J. A. & Little, R. J. A. (1995) *Biometrics*, in press.
4. Felsenstein, J. (1992) *Genet. Res.* **60,** 209–220.
5. Efron, B. (1979) *Ann. Stat.* **7,** 1–26.
6. Felsenstein, J. (1985) *Evolution* **39,** 783–791.
7. Bandelt, H.-J. & Dress, A. (1986) *Adv. Appl. Math.* **7,** 309–343.
8. Fitch, W. (1977) *Am. Nat.* **111,** 223–257.
9. Lake, J. A. (1991) *Mol. Biol. Evol.* **8,** 378–385.
10. Maslov, D. A., Avila, H. A., Lake, J. A. & Simpson, L. (1994) *Nature (London)* **368,** 345–348.
11. Halanych, K. M., Bacheller, J. D., Aguinaldo, A. A., Hillis, D. M. & Lake, J. A. (1995) *Science* **267,** 1641–1643.
12. Rivera, M. C. & Lake, J. A. (1992) *Science* **257,** 74–76.
13. Felsenstein, J. (1978) *Syst. Zool.* **27,** 401–410.
14. Mindel, D. P. (1991) in *Phylogenetic Analysis of DNA Sequences*, eds. Miyamoto, M. & Cracraft, J. (Oxford Univ. Press, Oxford), pp. 119–136.
15. Lake, J. A. (1988) *Nature (London)* **331,** 184–186.
16. Steel, M. A., Lockhart, P. J. & Penny, D. (1993) *Nature (London)* **364,** 440–442.
17. Eldridge, N. & Cracraft, J. (1980) *Phylogenetic Patterns and the Evolutionary Process* (Columbia Univ. Press, New York).
18. Woese, C. R. (1987) *Microbiol. Rev.* **51,** 221–271.
19. Jukes, T. H. & Cantor, C. R. (1969) in *Mammalian Protein Metabolism*, ed. Munro, H. N. (Academic, New York), Vol. 3, pp. 21–132.
20. Kimura, M. (1983) *The Neutral Theory* (Cambridge Univ. Press, Cambridge, U.K.).
21. Lake, J. A. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 1455–1459.
22. Lockhart, P. J., Steel, M. A., Hendy, M. D. & Penny, D. (1994) *Mol. Biol. Evol.* **11,** 605–615.
23. Steel, M. A. (1994) *Appl. Math. Lett.* **7,** 19–24.
24. Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., Manolson, M. F., Poole, R. J., Date, T., Oshima, T., Konishi, J., Denda, K. & Yoshida, M. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 6661–6665.
25. Iwabe, N., Kuma, K.-i., Hasegawa, M., Osawa, S. & Miyata, T. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 9355–9359.
26. Runnegar, B. (1993) *Early Life on Earth*, ed. Bengtson, S., (Cambridge Univ. Press, Cambridge, U.K.).
27. Lake, J. A. (1986) *Nature (London)* **319,** 626.
28. Cousineau, B., Cerpa, C., Lefebvre, J. & Cedergren, R. (1992) *Gene* **120,** 33–41.
29. Hasegawa, M., Hashimoto, T. & Adachi, J. (1992) in *The Origin and Evolution of Prokaryotic and Eukaryotic Cells*, eds. Hartman, H. & Matsuno, K. (World Scientific, Singapore), pp. 107–130.
30. Creti, R., Citarella, F., Tiboni, O., Sanangelantoni, A., Palm, P. & Cammarano, P. (1991) *J. Mol. Evol.* **33,** 332–342.
31. Cavender, J. A. (1978) *Math. Biosci.* **40,** 271–280.
32. Felsenstein, J. (1985) *Sys. Zool.* **34,** 152–161.
33. Churchill, G. A., von Haeseler, A. & Navidi, W. C. (1992) *Mol. Biol. Evol.* **6,** 753–769.
34. Buckley, F. & Harary, F. (1990) *Distance in Graphs* (Addison–Wesley, New York).